Managing the Data Deluge

Enhancing Big Data Analytics in Financial Services with Solid-State Storage
Managing the Data Deluge

Enhancing Big Data Analytics in Financial Services with Solid-State Storage

By Anne Rawland Gabriel

Across the industry, decision making is becoming data-driven. In fact, according to a recent survey by The Economist Intelligence Unit, 73 percent of C-level financial services leaders reported that more and more management decisions are based on “hard analytical information.”¹

Whether it’s meeting competitive, customer or compliance demands, financial services firms aren’t just striving to harness the proliferating volumes of historical data. Many organizations also want insights from new and unstructured sources such as news reporting, Web usage trends or social media chatter.

According to Gartner, “Business executives and IT managers are increasingly referring to information as one of their organization’s most critical and strategic corporate assets. Certainly there is a sharp rise in enterprises leveraging information as a performance fuel, competitive weaponry and a relationship adhesive.”²

Precisely what constitutes big data may be a moving target, but a single data set can range from dozens of terabytes to a data deluge encompassing multiple petabytes.

But the excitement in financial services isn’t about collecting and storing large amounts of data — a challenge the industry solved long ago. Instead, the buzz is about speeding the response times for data analysis and visualization and providing actionable information when it’s needed most.

To put it into perspective, let’s say your company monitors the trading activity of 10,000 global stocks. When a new crisis breaks out in an oil-producing region, your business users want to know what would happen — across the entire 10,000-stock landscape — if there were a 10 percent rise in the price of oil as a result of the disruption. Further, they want to model the effects of 20, 40 and 60 percent price increases.

For business users making billion-dollar decisions affecting their company and their clients,
getting immediate answers is imperative. And, they want to do so by running the analyses themselves, not by handing them off to a data expert in IT. In other words, for financial services, the issue isn’t the raw ability to manipulate large amounts of data; robust business intelligence (BI) tools have had that capability for some time. It’s the emergence of technologies — both hardware and software — that enable business users to visualize actual trends and perform what-if modeling quickly and efficiently.

“Although the business intelligence tools market exceeded $120 billion last year, our surveys show that only 30 percent of potential users actually use them,” notes Rita Sallam, research vice president at Gartner. The challenge, she says, is “having self-service BI tools appropriate for business users. A big reason for lack of adoption is that the tools are just too hard to use.”

Now, there’s a new breed of big-data applications, offering business users a dashboard-style interface that converts complex analytical calculations and models into digestible, intuitive and user-friendly charts and graphs. With a few keystrokes, users can not only create multivariate models but also view the effects of changing one or more variables.

“If you’re trying to answer an ad hoc question, not waiting a lengthy period for all the data to be formatted in your enterprise BI solution is a fantastic advancement,” says Jamon Bowen, principal technical marketing manager at LSI Corporation. For instance, consider the previously discussed crisis in the oil-producing region. With today’s big data tools, global events of all kinds — a natural disaster that shuts down a major tech manufacturing hub or a farmers’ strike in a major food-producing region in the Southern Hemisphere — can be analyzed in combination.

In addition to viewing how such events change an entire data set, many big data tools permit grouping inputs (such as company stocks) to get a sector or regional view. And, these apps allow users to drill down to see the impact on individual inputs.

Local Storage Makes It Happen
To manipulate data quickly enough to deliver the fast outputs required by financial services firms, many big data software applications minimize hardware-related delays by using local storage.

In a nutshell, today’s big data applications typically leverage multicore main memory CPUs in conjunction with direct-attached storage (DAS) in a server. This is faster than pulling data over a cable that extends to shared storage subsystems, whether network-attached storage (NAS) or a storage area network (SAN).

“Many of the big data structures are based on bringing processing to data rather than data to processing,” affirms Bowen. “So the trend in big data is scaling out using more locally attached storage.”

Indeed, many industry experts, like LSI’s Bowen, predict that tremendous gains in performance will only accelerate. “We’re coming into a golden age of parallel computing,” he says. “Big data frameworks help bring parallel computing out of the specialized domain of high-performance computing and into the typical enterprise.”

“And,” he continues, “with 16-core processors, hyper-threading and 10 Gigabit Ethernet networking, the numbers will just keep marching up.”

In a nutshell, this miniaturization and commoditization of grid computing makes it a competitive imperative for financial services organizations to take advantage of today’s massively parallel systems.

The Downside of Local Storage for Big Data
As is true for many new technologies, there can be a catch to using local storage for big data applications: hard disk drive-related performance bottlenecks.

In a typical big data architecture, each server in big data is referred to as a “node.” Multiple nodes are required and are arranged in clusters. Within a node (server), there are typically 12 to 24 high-performance enterprise hard disk drives (HDDs). And, although HDD performance metrics continue to climb, input/output (I/O) bottlenecks are a continuous reality.

Such bottlenecks become a challenge in big data because thousands of nodes (servers) are commonly used. All of the nodes work on the same data set simultaneously by performing functions and then providing the results to other nodes or other clusters of nodes for further processing.

From a technical standpoint, the root of the

4. “Big Results from Big Data,” Wall Street & Technology magazine, May 2012
issue occurs during the “MapReduce” operation, which consists of three different operations. The “Map” and “Reduce” functions are designed to stream data directly from disks. Between these two operations, data is passed between the Map and Reduce functions during a “shuffle” phase.

During this handoff, data passes between Map and Reduce by first using disk storage, and then transferring the data over the network. Given the thousands of nodes involved, even the smallest disk and network-related bottlenecks quickly add up to degrade output.

While increasing the performance of both the Map and Reduce operations has been the focus of those who created big data computing, less attention — thus far — has targeted performance degradation during the shuffle phase.

“The Map and Reduce functions have been tuned for bringing processing to data,” explains Bowen. “However, during the shuffle phase, the effects of passing data from thousands of nodes, to thousands of nodes, are frequently underestimated.”

“With so many nodes passing data, the I/O bottlenecks can be substantial,” he adds.

**Enter Flash-Based Application Acceleration Technologies**

Fortunately, financial services organizations can now address the performance degradation that occurs during the MapReduce shuffle phase with flash-based application acceleration solutions such as the LSI® Nytro™ product family.

Not surprisingly, adding sophisticated caching algorithms, such as those found in the Nytro application acceleration architecture, to a flash hardware card can take performance to the next level.

According to independent lab validation testing provider Demartek, “caching on flash … places a copy of ‘hot’ data into its cache so that I/O activity can be accelerated. This caching technique benefits any application whose data is considered ‘hot’ and is within the scope of management and visibility of the cache.”

Frequently called “server-side caching,” the technology accelerates I/O activity without requiring changes to the applications running in the host server, says Demartek. Among other things, this improves overall storage performance, including DAS-based storage, and drives down latency.

While performance gains vary depending on the situation, Bowen says using LSI Nytro application acceleration architecture can cut big data job runtimes by up to as much as 40 percent.

**Meeting Big Data Application Acceleration Needs**

Clearly, flash-based solutions, such as the LSI Nytro Application Acceleration Architecture, can help financial services users improve big data modeling capabilities to enable fast, compliant real-time decision making.

Specifically, the Nytro product family of intelligent solid-state storage includes two solutions appropriate for big data platforms in the financial services:

- **Nytro MegaRAID®** application acceleration cards are appropriate for off-the-shelf big data setups, where data placement during MapReduce is typically not addressed. In these circumstances, LSI Nytro MegaRAID intelligently places data behind the scenes without requiring any investment in big data application development. Additionally, the product provides data protection within big data nodes with LSI RAID-on-Chip (ROC) technology, which supplies RAID levels 0, 2, 5 and 6.

- **Nytro WarpDrive™** application acceleration cards are best for big data applications where the framework decides how to place data. Typically, this occurs in very large big data setups where application developers control data placement. Generally, there are known locations where a flash-based solution will improve performance most effectively.

Solutions such as the Nytro product family combine advanced software caching routines with flash memory hardware mounted on a PCI card. Since PCI cards easily slide into available server slots, no additional power, cooling or space demands are made.

Because flash storage offers dramatically lower data access times than disks, it can provide a superior platform for caching data. According to the Storage Networking Industry Association, the storage industry’s standards body, flash-based storage access times range from 100 to 1,000 times faster than for a mechanical HDD.

---

5 LSI® Nytro™ XD Caching Solution Evaluation, Demartek, LLC, April 2012.
6 Ibid.
No matter which solution is best for your organization, the LSI Nytro product family offers the following key features to benefit all types of big data environments:

- **Automatic Data Placement** detects data “hot spots,” intelligently determines the best location for that data and then places it onto the appropriate disk — all in the background and at lightning speed. This capability can significantly boost performance during the MapReduce operation.

- **Flexible Solution** permits deploying LSI Nytro PCIe cards with a wide variety of hardware and software, regardless of the customizations an organization has made to applications or operating systems. “For example, organizations frequently tweak the Linux® operating system,” Bowen points out. “By providing the driver source code with some Nytro solutions, flash-based application acceleration can be realized in these environments.”

- **Cost-Effective** to deploy within big data environments of any size. “It’s possible to use a very small amount of flash in comparison to hard disk,” says Bowen. “For example, 500GB of Nytro cards can accelerate up to 10TB of hard disk.”

- **Scalable** to meet the needs of expanding big data demands. Nytro PCI cards offer various scalability options. A single Nytro MegaRAID supports up to 128 hard disk drives within a big data node. And, Nytro WarpDrive comes in capacity options ranging from 200GB to 3.2TB.

In short, flash-based application acceleration can provide financial services organizations with the missing piece in the big data puzzle. By addressing inherent hard disk-related latencies during the MapReduce operation, the LSI Nytro product family is playing a significant role in delivering the real-time analytics that organizations need to meet competitive, customer and compliance demands.

---

**ABOUT LSI**

LSI Corporation (NYSE: LSI) designs semiconductors and software that accelerate storage and networking in data centers, mobile networks and client computing. Our technology is the intelligence critical to enhanced application performance and is applied in solutions created in collaboration with our partners. More information is available at [www.lsi.com](http://www.lsi.com).

LSI, the LSI & Design logo, MegaRAID, Nytro and WarpDrive are trademarks or registered trademarks of LSI Corporation. All other brand or product names may be trademarks or registered trademarks of their respective companies.